# Real-time Quadrifocal Visual Odometry

A.I. Comport* , E. Malis† and P. Rives†

*Abstract*—

**This paper describes a new image-based approach to tracking the 6 degrees of freedom trajectory of a stereo camera pair using a corresponding reference image pair whilst simultaneously determining pixel matching between consecutive images in a sequence. A dense minimisation approach is employed which directly uses all grey-scale information available within the stereo pair (or stereo region) leading to very robust and precise results. Metric 3D structure constraints are imposed by consistently warping corresponding stereo images to generate novel viewpoints at each stereo acquisition. An iterative non-linear trajectory estimation approach is formulated based on a quadrifocal relationship between the image intensities within adjacent views of the stereo pair. A robust M-estimation technique is used to reject outliers corresponding to moving objects within the scene or other outliers such as occlusions and illumination changes. The technique is applied to recovering the trajectory of a moving vehicle in long and difficult sequences of images.**

## I. INTRODUCTION

Here the core issue of 3D visual odometry is considered in the context of rapidly moving vehicles, real sequences, large scale distances, with traffic and other types of occluding information. Indeed, tracking in urban canyons is a non-trivial problem [2], [1]. It is clear that pose estimation and visual tracking are also important in many applications including robotic vision, augmented reality, medical imaging, etc...

Model-based techniques have shown that 3D CAD models are essential for robust, accurate and efficient 3D motion estimation [3], however, they have the major drawback of requiring an a-priori model which is not always available or extremely difficult to obtain as in the case of shapeless objects or large-scale environments.

Alternative techniques propose to perform 3D structure and motion estimation online. Among this class, visual simultaneous localisation and mapping approaches [4], [5] are classically based on an implementation of the Extended Kalman Filter and have limited computational efficiency (manipulation and inversion of large feature co-variance matrices) and limited inter-frame movements (due to approximate non-iterative estimation). In [6] stereo and monocular visual odometry approaches are proposed based on a combination of feature extraction, matching, tracking, triangulation, RANSAC pose estimation and iterative refinement. In [2], a similar monocular technique is proposed but drift is minimised using a local bundle adjustment technique.

Stereo techniques provide accurate 3D information at little computational cost (1D search and matching along epipolar lines) and subsequently avoid the problems of monocular algorithms (i.e. scale factor, initialisation, observability, etc.) by using prior knowledge about the extrinsic camera parameters and applying multi-view constraints. Indeed a multitude of work exist on multiview-geometry (see [7] and ref. therein). State of the art stereo techniques [8], [9] are, however, only feature based and to our knowledge no work has been done on deriving an efficient direct tracker as in [10], [11] using stereo warping and novel view synthesis as in [12].

Feature based methods [4], [5], [6], [2], [13] all rely on an intermediary estimation process based on detection thresholds. This feature extraction process is often badly conditioned, noisy and not robust therefore relying on higher level robust estimation techniques. Since the global estimation loop is never closed on the image measurements (intensities) these multi-step techniques systematically propagate feature extraction and matching errors and accumulate drift (refer to Figure 1(a)). It is important to note that stereo feature based techniques require spatial matching across the stereo pair and temporal matching between stereo pairs. To eliminate drift these approaches resort to techniques such as local bundle adjustment or SLAM. On the other hand, feature based techniques have the advantage of allowing matching across the entire image and therefore can handle large inter-frame movements. Unfortunately, if there is little overlap between images, feature based approaches are nonetheless prone to miss-matching and are unstable since there are fewer features to be matched and the esti-

* The first author is with the CNRS and carried this work between LASMEA, Clermont-Ferrand and INRIA Sophia-Antipolis. comport@i3s.unice.fr

† The second two authors are with INRIA, Sophia-Antipolis, France. name.surname@sophia.inria.fr

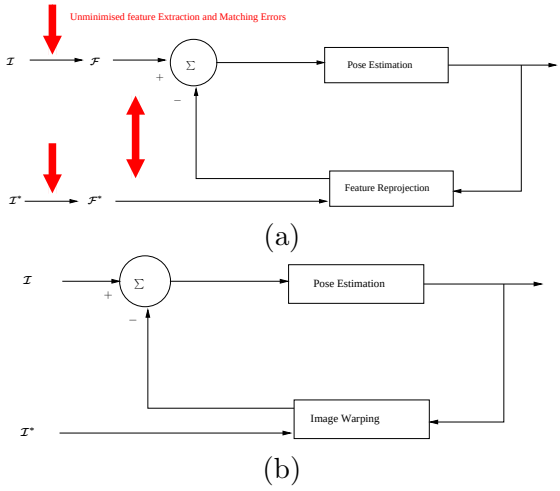



Fig. 1. Two non-linear iterative estimation loops. (a) A feature based approach where it can seen that there unminimized errors made by the extraction of stereo features $\mathcal{F}$ from both the current stereo images $\mathcal{I}$ and the reference stereo images $\mathcal{I}^*$. Furthermore, there are unminimized errors from the matching phase between the stereo reference features and the current ones. (b) A direct minimisation approach that minimises directly the pixel intensities between a warped image and the reference image. Here the errors shown in (a) are minimised. Note that in this paper, only the temporal correspondence errors are minimised and not the spatial ones. This is left as a perspective of the approach.

mator can become ill-conditioned.

Appearance, optical flow or direct techniques [14], [15], on the other hand, are image-based and minimise an error directly based on the image measurements (refer to Figure 1(b)). These approaches have the advantage of being precise and perform tracking and pixel correspondence/matching simultaneously. Unfortunately, techniques published so far have only considered tracking between two monocular images (including spatial stereo matching) and often make heavy assumptions about the nature of the structure within the scene or the camera model. For example in [16] an affine motion model is assumed and in [17] and [10] planar homography models are assumed. In this way the perspective effects or the effects of non-planar 3D objects are not considered and tracking fails easily under large movements. Furthermore, these techniques all require the definition of a region of interest within the image and are limited to local convergence around that region. More specifically they require a *sufficient overlap* as opposed to feature based techniques which can perform matching globally within the image.

Another very important issue is the registration problem. *Purely geometric*, or *numerical and iterative* approaches may be considered. *Linear approaches* use a least-squares method to estimate the pose and are considered to be more suitable for initialization procedures. *Full-scale non-linear optimisation techniques* (e.g. [16], [17], [3]) consist of minimizing an objective function using numerical iterative algorithms such as Newton-Raphson or Levenberg-Marquardt. The main advantage of these approaches are their computational efficiency and their accuracy, however, they may be subject to local minima and, worse, divergence.

The technique proposed in this paper is a 3D visual odometry technique that minimises a direct intensity error between consecutive stereo pairs. This approach lies at the intersection between direct image-based and model based techniques. The image-based model is obtained by performing dense stereo matching either online when in unknown environments, or off-line when a training set is available. The 3D model then comprises both photometric stereo image information along with a disparity map. The global image-based model can then be considered as a set of key reference image-pairs that are used to perform localisation locally around those reference positions. Six degrees of freedom pose estimation is achieved by defining a quadrifocal warping function which closes a non-linear iterative estimation loop *directly* with the images. This approach handles arbitrary 3D structure and improves the convergence domain with respect to region-based methods since the entire image is used and the probability of a sufficient inter-frame overlap is therefore much higher. Whilst this approach improves the convergence domain, the accuracy of direct techniques [15] is retained since no feature extraction is performed. In terms of minimisation, in this paper an efficient second-order approach [18], [10] is employed which improves efficiency and also helps to avoid local minima.

As will be shown, the proposed technique is able to accurately handle large scale scenes efficiently whilst avoiding error prone feature extraction and inter-frame matching. This leads to very impressive results in real-scenes with occlusions, large inter-frame displacements, and very little drift. This paper is an extended and more detailed version of the technique presented in [19]. In Section II an overview of the objective function is given. Section III-B the stereo warping function is detailed. Section IV outlines the robust second order minimisation technique and in VI the results are presented.

## II. Trajectory Estimation

A framework is described for estimating the trajectory of a stereo-camera rig along a sequence from a designated region within the image. The tracking problem will essentially be considered as a pose estimation problem which will be related directly to the grey-level brightness measurements within the stereo pair via a non-linear model which accounts for the 3D geometric configuration of the scene.

Since the ultimate objective is to control a robot within Euclidean space, a calibrated camera pair is considered. Consider a stereo camera pair with two brightness functions $\mathbf{I}(\mathbf{p},t)$ and $\mathbf{I}'(\mathbf{p}',t)$ for the left and right cameras respectively, where $\mathbf{p} = (u,v,1)$ and $\mathbf{p}' = (u',v',1)$ are homogeneous vectors containing the pixel locations within the two images acquired at time $t$. It is convenient to consider the set of image measurements in vector form such that $\boldsymbol{\mathcal{I}} = (\mathbf{I},\mathbf{I}')^\top \in \mathbb{R}^{2n}$ is a vector of intensities of the left image stacked on top of the right, with $n$ the number of pixels in one image.

$\boldsymbol{\mathcal{I}}$ will be called the *current* view pair and $\boldsymbol{\mathcal{I}}^*$ as the *reference* view pair. A superscript $*$ will be used throughout to designate the reference view variables. Similarly, with abuse of notation, $\boldsymbol{\mathcal{P}}^* = (\mathbf{p}^*,\mathbf{p}^{*\prime}) \in \mathbb{R}^4$, is a stereo image correspondence from the reference template pair. Any set of corresponding pixels from the reference image-pair are considered as a reference template, denoted by $\mathcal{R}^* = \{\{\mathbf{p}^*{}_1,\mathbf{p}^{*\prime}{}_1\},\{\mathbf{p}^*{}_2,\mathbf{p}^{*\prime}{}_2\},\ldots,\{\mathbf{p}^*{}_n,\mathbf{p}^{*\prime}{}_n\}\}$ where $n$ is the number of corresponding point pairs in the template.

The motion of the camera pair or objects within the scene induces a deformation of the reference template. The 3D geometric deformation of a stereo rig can be fully defined by a motion model $w(\boldsymbol{\mathcal{P}}^*,\mathbf{T}',\mathbf{K},\mathbf{K}';\overline{\mathbf{T}}(t))$. The motion model $w$ considered in this paper is the quadrifocal warping function which will be detailed further in Section III. $\mathbf{K}$ and $\mathbf{K}'$ contain the intrinsic calibration parameters for the left and right cameras respectively. $\mathbf{T}' = (\mathbf{R}',\mathbf{t}') \in \mathbb{SE}(3)$ is the homogeneous matrix of the extrinsic camera pose of the right camera w.r.t. the left and $\overline{\mathbf{T}} = (\bar{\mathbf{R}},\bar{\mathbf{t}}) \in \mathbb{SE}(3)$ is the current pose of the stereo rig relative to the reference position. Throughout, $\mathbf{R}$ is a rotation matrix and $\mathbf{t}$ the translation vector. Since both the intrinsic and extrinsic calibration parameters do not vary with time they will be assumed implicit.

It follows that the reference image is obtained by warping the current image as:

$$\boldsymbol{\mathcal{I}}^*(\boldsymbol{\mathcal{P}}^*) = \boldsymbol{\mathcal{I}}\big(w(\boldsymbol{\mathcal{P}}^*;\overline{\mathbf{T}})\big), \quad \forall \boldsymbol{\mathcal{P}}^* \in \mathcal{R}^*. \tag{1}$$

where $\overline{\mathbf{T}}$ is the true pose. When the coordinates indexing the image do not correspond to an exact pixel location bi-linear interpolation is performed.

Suppose that at the current image an estimate of the pose $\widehat{\mathbf{T}}$ fully represents the pose of the stereo pair with respect to a pair of reference images. The tracking problem then becomes one of estimating the incremental pose $\mathbf{T}(\mathbf{x})$, where $\mathbf{x}$ is a minimal parametrisation of the homogeneous pose matrix $\mathbf{T}$ and where it is supposed that $\exists\tilde{\mathbf{x}} : \widehat{\mathbf{T}}\mathbf{T}(\tilde{\mathbf{x}}) = \overline{\mathbf{T}}$. The estimate is updated by a homogeneous transformation $\widehat{\mathbf{T}} \leftarrow \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})$. It can be noted here that the increment is now parametrized to the right of $\widehat{\mathbf{T}}$ as opposed to [19] where it is parametrized to the left ($\mathbf{T}_L(\mathbf{x})$). Both cases are related by:

$$\mathbf{T}(\mathbf{x}) = \widehat{\mathbf{T}}^{-1}\mathbf{T}_L(\mathbf{x})\widehat{\mathbf{T}}. \tag{2}$$

This allows to simplify the derivation of the Jacobian (described in Appendix IX-A) so that a pre-calculation of the Jacobian can be made to improve the computational efficiency.

The unknown parameters $\mathbf{x} \in \mathbb{R}^6$ are defined by the Lie algebra $\mathfrak{se}$ as:

$$\mathbf{x} = (\boldsymbol{\omega}\Delta t, \boldsymbol{\upsilon}\Delta t) \in \mathfrak{se}, \tag{3}$$

which is the integral of a constant velocity twist which produces a pose $\mathbf{T}$, where $\boldsymbol{\upsilon}$ and $\boldsymbol{\omega}$ are the linear and angular velocities respectively. The pose and the twist are related via the exponential map as:

$$\mathbf{T}(\mathbf{x}) = exp\left(\begin{bmatrix} [\boldsymbol{\omega}]_\times & \boldsymbol{\upsilon} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right), \tag{4}$$

where $[.]_\times$ represents the skew symmetric matrix operator.

Thus the pose and the trajectory of the camera pair can be estimated by minimising a non-linear least squares cost function:

$$C(\mathbf{x}) = \sum_{\boldsymbol{\mathcal{P}}^* \in \mathcal{R}^*} \left(\boldsymbol{\mathcal{I}}\left(w(\boldsymbol{\mathcal{P}}^*;\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})\right) - \boldsymbol{\mathcal{I}}^*\left(\boldsymbol{\mathcal{P}}^*\right)\right)^2. \tag{5}$$

This function is minimised using the robust, efficient and precise second order minimisation procedure detailed in Section IV.

## III. Novel view synthesis and warping

The geometric configuration of a stereo pair, that is undergoing movement within a rigid scene, is based on the paradigm that four views of a scene satisfy quadrifocal constraints. Thus given a reference stereo view with correspondences between pixels and the quadrifocal tensor, a third view and fourth view can be generated by means of a warping function. This warping function subsequently provides the required relationship between two views of the scene and an adjacent pair of views in a sequence of images.

The approach presented here is formalised using the quadrifocal tensor since it encapsulates all the geometric relations between four views that are independent of scene structure and provides a clear insight into the geometric properties (homography transfer between two views via a line in the third, point-line relations, ...). This allows to clearly define a choice of quadrilinear constraints from the full set of possible constraints and to identify any degenerate configurations. In this way a clear link is made with projective multi-view geometry and extensions to our approach are therefore facilitated (for example, the extension to a fully projective approach). Furthermore, the analytical development of the image warping function ensures that the measurement uncertainties are consistently handled in the optimisation procedure developed in Section IV.

### A. Quadrifocal Geometry

A point $\mathbf{X} \in \mathbb{R}^3$ in 3D Euclidean space projects onto the 3D camera plane by a $3 \times 4$ projection matrix $\mathbf{M} = \mathbf{K}[\mathbf{R}|\mathbf{t}] \in \mathbb{P}^3$ where the image point is given by $\mathbf{p} = \mathbf{M}\mathbf{X}$ so that $\mathbf{p} = (u, v, 1)^\top$ is the homogeneous pixel vector (see Figure 2). It is assumed that the stereo-rig is calibrated with intrinsic camera parameters $\mathbf{K}$ and $\mathbf{K}'$ for the left and right cameras respectively and extrinsic parameters $\mathbf{T}'$ denoting the pose from the left to right camera.

Much work has been carried out on multi-view geometry and of particular interest here are the quadrilinear relations [20], [21], [22], [23], [24] between two pairs of stereo images at two consecutive time instants. The compact tensor notation of multifocal geometry will be used here with a covariant-contravariant summation convention. Contravariant point vectors $\mathbf{p}^i$ are denoted with a superscript and their covariant counterpart representing lines $\mathbf{l}_j \in \mathbb{P}^2$, are denoted with a subscript. A contraction or summation over two tensors occurs when there are repeated indices in both contravariant and covariant variables (i.e. $\mathbf{p}^i\mathbf{l}_i = \sum_{j=1}^{n} \mathbf{p}^j\mathbf{l}_j$). An outer-product of two first order tensors (vectors), $\mathbf{a}_i\mathbf{b}^j$ is a second order tensor (matrix) $\mathbf{c}_i^j$ which is equivalent to $\mathbf{C} = \mathbf{b}\mathbf{a}^\top$ in matrix notation.

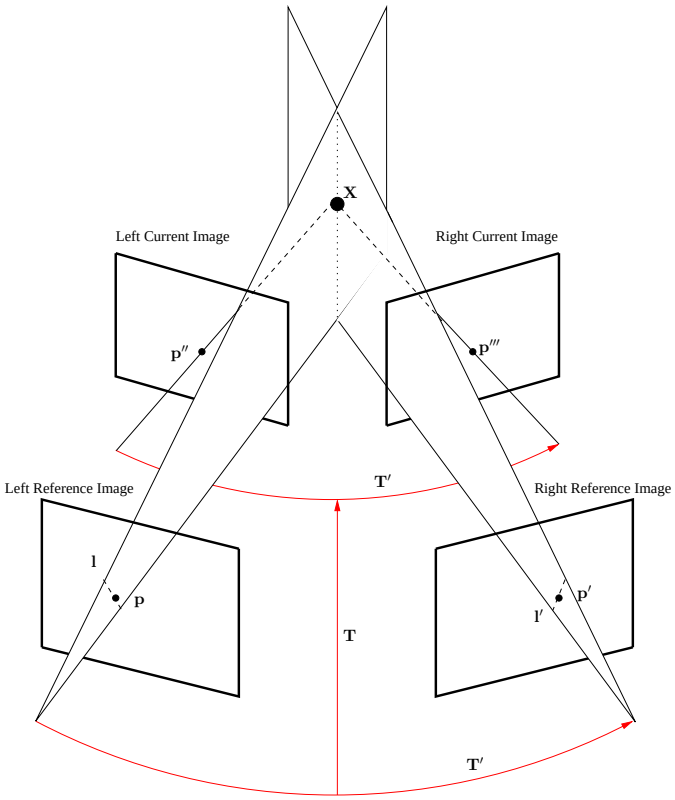



Fig. 2. The quadrifocal geometry of a stereo pair at two subsequent time instants. Two points $\mathbf{p}$ and $\mathbf{p}'$ are initialised only once at the beginning of the tracking process to be in correspondence. The central pose $\mathbf{T}$ is estimated via a non-linear warping function which warps all points in the reference stereo pair to the current image points $\mathbf{p}''$ and $\mathbf{p}'''$. The quadrifocal warping function is defined by choosing two lines $\mathbf{l}$ and $\mathbf{l}'$ passing through corresponding points in the first image. The extrinsic parameters $\mathbf{T}'$ are assumed known a-priori.

Consider a point in correspondence across four views, $\mathbf{p} \leftrightarrow \mathbf{p}' \leftrightarrow \mathbf{p}'' \leftrightarrow \mathbf{p}'''$, with the camera matrices $\mathbf{M}, \mathbf{M}', \mathbf{M}'', \mathbf{M}'''$. A common method for deriving the quadrifocal tensor is then to define the linear system as

$$\begin{bmatrix} \mathbf{M} & \mathbf{p} & & & \\ \mathbf{M}' & & \mathbf{p}' & & \\ \mathbf{M}'' & & & \mathbf{p}'' & \\ \mathbf{M}''' & & & & \mathbf{p}''' \end{bmatrix} \begin{pmatrix} \mathbf{X} \\ -k \\ -k' \\ -k'' \\ -k''' \end{pmatrix} = \mathbf{0} \qquad (6)$$

where $k, k', k'', k'''$ are unknown scale factors.

The matrix forming the left hand part of expres-

sion (6) is of dimension $12 \times 8$. Due to the existence of a solution and the fact that the right term is not zero, any $8\times8$ minor has a zero determinant. This fact defines the quadrilinear relations between the points in the four views. From the choice of the 8 rows from the various camera matrices two different cases may be considered:

**1**. The case where 8 rows are selected such that only one row from a camera matrix is included in the determinant leads to a trilinear or bilinear relationship. This case is not considered here.

**2**. The case where two rows from each camera matrix is included leads to a quadrifocal relationship, considered here.

The quadrifocal tensor that transfers a point in the left reference view, that is in correspondence with a point in the right reference view, to points in the current left and right views is then written as:

$$\mathbf{p}''^{p}\mathbf{p}'''^{q} = \mathbf{p}^{i}\mathbf{p}'^{j}\epsilon_{ipw}\epsilon_{jqx}\mathbf{Q}^{pqrs} \tag{7}$$

where $\epsilon$ is the permutation tensor with properties that can be used to represent the vector cross product, the skew symmetric matrix and the determinant. It is provided here for completeness:

$$\epsilon_{ijk}=\begin{cases} 0 & \text{unless } i,j,k \text{ are distinct} \\ +1 & \text{if } i,j,k \text{ are an even permutation of } 1,2,3 \\ -1 & \text{if } i,j,k \text{ are an odd permutation of } 1,2,3 \end{cases} \tag{8}$$

The quadrifocal tensor is a fourth order tensor represented by a homogeneous $3 \times 3 \times 3 \times 3$ array of elements. It is obtained by expanding the determinant of the matrix in (6) in terms of the points $\mathbf{p}, .., \mathbf{p}'''$ as:

$$\mathbf{Q}^{pqrs} = \det \begin{bmatrix} \mathbf{m}^{p} \\ \mathbf{m}'^{q} \\ \mathbf{m}''^{r} \\ \mathbf{m}'''^{s} \end{bmatrix} \tag{9}$$

where the contravariant indices $p, q, r, s$ index the rows of the camera matrices.

In order to transform the points in the reference images to points in the current images it is possible to either use a single quadrifocal tensor or to define two quadrifocal tensors, one for each reference image. This choice depends on the stereo matching precision (sub-pixel or one-to-one). First define the left and right quadrifocal tensors as $\mathbf{Q}_L$ and $\mathbf{Q}_R$). These two cases are then:

**1**. Singe quadrifocal tensor - In this case it is necessary to perform one-to-one matching between the reference images so that $\mathbf{Q}_L = \mathbf{Q}_R$. The disadvantage here is that the correspondences between left and right reference are only approximate and to not have sub-pixel matching accuracy. Furthermore, most one-to-one dense correspondence algorithms do not give the same results in each direction. The advantage is that it is computationally twice as fast as case 2.

**2**. Two quadrifocal tensors - In this case it is possible to perform sub-pixel matching for both left and right images , however, there are twice as many constraints to be applied.

These two cases are also a result of the decision to perform *direct* minimisation with the image measurements. It can be noted, however, that if the estimation loop is not closed *directly* with both images, then it is possible to perform sub-pixel matching with the left image as the base (for instance) and to approximate the right image by interpolating the corresponding pixels so as to maintain a one-to-one mapping. In this case the computation is efficient, sub-pixel matching is performed, however, the estimation is biased towards one image (i.e. the error being minimised is only direct wrt. one image). This approximation can subsequently lead to inaccurate results whether it be inaccurate correspondences or a bias with respect to errors made in the intrinsic parameters of a 'dominant' camera.

In this paper the dual case will be developed and the following paragraphs will show how this relationship can be simplified into a single quadrifocal tensor whilst maintaining symmetry. In [24] a detailed method for decomposing the quadrifocal tensor as an epipole-homography pairing is given. Of interest here is the composition of the quadrifocal tensor from two trifocal and a bifocal relationship.

$$\delta_i\mu_j\mathbf{Q}^{ijkl} = [\delta_i\mu_j\mathcal{T}_l^{ij}]_x\mathbf{F}_{12}[\delta_i\mu_j\mathcal{T}_k^{ij}]_x \tag{10}$$

where $x$ is an index from the set $i, j, k, l$ and $\delta_i$ and $\mu_j$ range over the standard bases $(1,0,0), (0,1,0), (0,0,1)$. Under this definition every $3\times3\times3$ slice of $\mathbf{Q}^{ijkl}$ corresponds to Homography tensor [24] between the remaining views not represented by $x$.

Following from the previous discussion, each of the two sub-pixel quadrifocal tensors provides a relationship between corresponding points in the four images that can be decomposed into two trifocal tensors and a fundamental matrix. This decomposition will be used here to reduce the number of quadrilinear relations to a single set, whilst maintaining the one-to-many correspondences for both the left and right reference

images. With the given decomposition (and before simplification), this makes a total of two fundamental matrices and 4 trifocal tensors:
**1**. Two fundamental matrices between the left and right reference images (1 in each direction): Since the pose between the reference and current viewpoints is initially unknown then the transfer of points between the left and right reference views is a bilinear relation that depends on the matched points between the two images. It is assumed here that the matching is performed a-priori by a dense correspondence technique that exploits this bilinear relation via a 1D search along epipolar lines.
**2**. Two trifocal tensors relating each reference image to its current image via a corresponding reference image: This is the most important case as it depends on the unknown pose and remains robust to camera modeling errors.
**3**. Two trifocal tensors relating each reference image to the opposite current image: This case is interesting since it depends on the unknown pose, however, this case is more sensitive to camera calibration or modeling errors. This could be a good candidate if one wished to estimate the calibration parameters.

In order to reduce both quadrifocal tensors into a single relation, the left-right fundamental matrix is retained along with the two trifocal tensors given by case 2 above. In this way the quadrilinear constraints are maintained while reducing the computational complexity of the optimisation. Thus from (10), $\mathcal{T}_l^{ij}$ is chosen as the trifocal tensor between views (4,1,2), $\mathcal{T}_k^{ij}$ is chosen as the trifocal tensor between views (3,1,2) and $\mathbf{F}_{12}$ the fundamental matrix between views (1,2).

The geometry between two stereo pairs is therefore defined in a manner that is simple for subsequent developments using the canonical coordinates of two triplets of images. First of all, consider the triplet consisting of the left reference camera, the right reference camera and the left current camera. The left reference camera matrix is chosen as the origin so that $\mathbf{M} = \mathbf{K}[\mathbf{I}|\mathbf{0}]$. The reference projection matrix for the right camera (the extrinsic camera pose) and the *current* projection matrices for the left camera are then as : $\mathbf{M}' = \mathbf{K}'[\mathbf{R}'|\mathbf{t}']$ and $\mathbf{M}'' = \mathbf{K}[\mathbf{R}''|\mathbf{t}'']$.

The second triplet is defined in a similar manner such that the right reference camera is chosen as the origin and the left reference camera and right current camera matrices are defined with respect to this origin.

In order to construct the quadrifocal relation it is necessary to combine these two triplets of images using the left-right bilinear relation. This is done symmetrically by defining the arbitrary *world origin* as the geodesic center between the two reference cameras. To do this the extrinsic parameters are separated into two distinct poses with respect to the center as:

$$\mathbf{T}^{\mathbf{c}} = exp^{(\log(\mathbf{T}'))/2} \quad \text{and} \quad \mathbf{T}^{\mathbf{c}\prime} = \mathbf{T}^{\mathbf{c}}\mathbf{T}'^{-1}, \tag{11}$$

where $e$ and log are the matrix exponential and logarithm.

The pose from the left reference camera to the current one is therefore composed of a central pose as:

$$\mathbf{T}'' = \mathbf{T}^{\mathbf{c}-1}\widetilde{\mathbf{T}}\mathbf{T}^{\mathbf{c}}, \tag{12}$$

where $\widetilde{\mathbf{T}}$ is the unknown pose to be estimated. It will be shown in Section IV how this pose may be estimated iteratively.

### *B. Quadrifocal warping*

The quadrifocal warping function $w(\boldsymbol{\mathcal{P}}^*; \overline{\mathbf{T}})$ from (5) can now be considered to be composed of a trifocal tensor for each of the left and right images, that depend on the unknown minimal pose parameters, along with the constant extrinsic camera pose that provides the bilinear constraint of equation (10). Since, the bilinear constraint corresponds to a constant change of reference frame, this will be performed during the estimation in Section IV. For the moment the focus will be made on defining the warping of each left and right image via their corresponding trifocal tensors. In overview, the trifocal tensor is used to transfer (warp) corresponding points from two views to a third view. This tensor depends only on the relative motion of the cameras as well as the intrinsic and extrinsic camera parameters.

The trifocal tensor $\mathcal{T}$ is a third order tensor represented by a homogeneous $3 \times 3 \times 3$ array of elements. The trifocal tensor can be determined from equation (9) by taking into account the case of one line of the last camera matrix. The calibrated case is given as:

$$\begin{aligned} \mathcal{T}_i^{jk} = \; & \mathbf{k}'^{j}_{m}\mathbf{r}'^{m}_{n}\mathbf{k}^{-1}{}^{n}_{i}.\mathbf{k}''^{k}_{o}\mathbf{t}''^{o}(t) \\ & -\mathbf{k}'^{j}_{p}\mathbf{t}'^{p}.\mathbf{k}''^{k}_{q}\mathbf{r}''^{q}_{r}(t)\mathbf{k}^{-1}{}^{r}_{i}, \end{aligned} \tag{13}$$

where $(\mathbf{r}', \mathbf{t}')$ and $(\mathbf{r}'', \mathbf{t}'')$ are the tensor forms of the rotation matrix and translation vector for the second and third camera matrices respectively. $\mathbf{k}$ and $\mathbf{k}'$ are the intrinsic calibration components of the left and right camera matrices respectively. Note that $\mathbf{k}'' = \mathbf{k}$

or $\mathbf{k}'' = \mathbf{k}'$ depending on whether one is warping to the left or right camera at the next time instant.

Given any line $\mathbf{l}$ coincident with $\mathbf{p}$ or any line $\mathbf{l}'$ coincident with $\mathbf{p}'$ then the trifocal tensor contracts so as to become a homography $\mathbf{h}$ which maps points from one reference image to the current image. i.e. a line defined in one of the reference views defines a plane which can be used to warp a point between the remaining reference image and the current image. Thus the warping from the left reference image to the left current image via a plane in the right reference image is given by:

$$\mathbf{p}''^k = \mathbf{p}^i \mathbf{l}'_j \mathcal{T}_i^{jk} = \mathbf{h}_i^k \mathbf{p}^i,$$

where $\mathbf{p}^i$ is a point in the left reference image, $\mathbf{l}_k$ is a line defined in the right reference image and $\mathbf{p}''^k$ is the warped point in the left current image. This equation is used similarly for warping a point in the right reference image to a point in the right current via a plane in the left reference image.

As opposed to transfer using the fundamental matrix, the tensor approach is free from singularities when the 3D point lies on the trifocal plane. The only degenerate situation that occurs is if a 3D point lies on the baseline joining the first and the second cameras since the rays through $\mathbf{p}$ and $\mathbf{p}'$ are co-linear.

It is, however, important to carefully choose the seemingly arbitrary lines passing through the points $\mathbf{p}$ and $\mathbf{p}'$. In particular, the trifocal tensor is not defined when the epipolar line is chosen. The lines may, however, be chosen in several ways [7] including an optimal least squares solution to the linear system of equations $\mathbf{p}^i(\mathbf{p}'^j\epsilon_{jpr})(\mathbf{p}''^k\epsilon_{kqs})\mathcal{T}_i^{pq} = \mathbf{0}_{rs}$, where $\epsilon$ is the tensor which transforms $\mathbf{p}$ into its skew-symmetric form $[\mathbf{p}]_\times$. This choice, however, immensely complicates the analytic derivation of the Jacobian given in section IV. Other alternatives include choosing the line perpendicular to the epipolar line, or computing the result using several lines and choosing the best one. In the case of a calibrated stereo-rig the epipolar geometry is known so it is possible to directly choose a single line that is not degenerate and at the same time one that simplifies the derivation of the Jacobian. Subsequently, the reference line is chosen to be the diagonal line $\mathbf{l} = (-1, -1, u + v)$ coincident with the point $(u, v)$.

The stereo warping operator is then given by:

$$\begin{bmatrix} \mathbf{p}''^k \\ \mathbf{p}'''^n \end{bmatrix} = \begin{bmatrix} \mathbf{p}^i \mathbf{l}'_j \mathcal{T}_i^{jk} \\ \mathbf{p}'^l \mathbf{l}_m \mathcal{T}_l^{mn} \end{bmatrix}, \tag{14}$$

where the indexes of the two trifocal-tensors indicate tensors transferring to the left and right cameras. The lines $\mathbf{l}'$ and $\mathbf{l}$ are chosen to be the diagonal line as outlined in the previous paragraph. If these trifocal tensors are contracted into constant $(\mathbf{p}, \mathbf{p}', \mathbf{l}, \mathbf{l}', \mathbf{r}', \mathbf{t}', \mathbf{k}, \mathbf{k}')$ and non-constant components $(\mathbf{r}'', \mathbf{t}'')$, the warping function is composed of projective 3D points (expanded with the current calibration parameters) and the unknown pose.

It is important for further developments to highlight that the warping operator $w(\boldsymbol{\mathcal{P}}^*; \overline{\mathbf{T}}) : \mathbb{SE}(3) \times \mathbb{R}^4 \longrightarrow \mathbb{R}^4$ is a *group action*. Indeed, the following operations hold:

**1**. The identity map:

$$w(\boldsymbol{\mathcal{P}}^*; \mathbf{I}) = \boldsymbol{\mathcal{P}}^*, \quad \forall \boldsymbol{\mathcal{P}}^* \in \mathbb{R}^4, \tag{15}$$

**2**. The composition of an action corresponds to the action of a composition $\forall \mathbf{T_1}, \mathbf{T_2} \in \mathbb{SE}(3)$:

$$\begin{array}{l} w(w(\boldsymbol{\mathcal{P}}^*, \mathbf{T_1}), \mathbf{T_2}) = \\ w(\boldsymbol{\mathcal{P}}^*, \mathbf{T_1}\mathbf{T_2}) \end{array} \qquad \forall \boldsymbol{\mathcal{P}}^* \in \mathbb{R}^4. \tag{16}$$

## IV. Robust and Efficient Quadrilinear Tracking

The aim now is to minimise the difference in image intensities from the objective criterion defined previously (5) in an accurate and robust manner. If the $L_2$ norm is chosen the approach is well known as sum-of-squared difference (SSD) tracking. If a robust objective function is considered then the objective function therefore becomes:

$$O(\mathbf{x}) = \sum_{\boldsymbol{\mathcal{P}}^* \in \mathcal{R}^*} \rho\left(\boldsymbol{\mathcal{I}}(w(\boldsymbol{\mathcal{P}}^*; \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})) - \boldsymbol{\mathcal{I}}^*(\boldsymbol{\mathcal{P}}^*)\right), \tag{17}$$

where $\rho(u)$ is a robust function [25] that grows subquadratically and is monotonically non-decreasing with increasing $|u|$ (see Appendix II).

Since this is a non-linear function of the unknown pose parameters an iterative minimisation procedure is employed. The robust objective function is minimized by: $[\nabla O(\mathbf{x})]_{\mathbf{x}=\tilde{\mathbf{x}}} = 0$, where $\nabla$ is the gradient operator with respect to the unknown parameters (3) and there exists a stationary point $\mathbf{x} = \tilde{\mathbf{x}}$ which is the global minimum of the cost function.

Pseudo second order methods such as Levenberg-Marquardt are generally employed to minimise iteratively such an objective function since computing the Hessian to obtain full quadratic convergence is computationally expensive. However, since both the reference image and current image are available, along

with the fact that the warping operator has group properties, it is possible to use the efficient second order approximation (ESM) proposed in [18], [10]. This technique allows to avoid the computation of the Hessian.

For completeness the essential steps are summarised here. Consider the general least-squares minimisation problem:

$$F(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{n}(f_i(\mathbf{x}))^2 = \frac{1}{2}\|\mathbf{f}(\mathbf{x})\|^2 \tag{18}$$

In order to minimise this non-linear function an iterative gradient descent is performed in order to search for $\nabla F(\tilde{\mathbf{x}}) = \mathbf{0}$, where the definition of $\nabla$ is taken to be the multi-variate, multi-function Jacobian gradient operator which takes the derivative of $F(\mathbf{x})$ with respect to $\mathbf{x}$. To derive the ESM, the first step is a second-order Taylor series expansion of $\mathbf{f}$ about $\mathbf{x} = \mathbf{a}$ as:

$$\begin{aligned}\mathbf{f}_i(\mathbf{x}) = \; & \mathbf{f}_i(\mathbf{a}) + \nabla_i^j \mathbf{f}(\mathbf{a})(\mathbf{x}_j - \mathbf{a}_j) \\ & + \tfrac{1}{2}\nabla_i^{jk}\mathbf{f}(\mathbf{a})(\mathbf{x}_j - \mathbf{a}_j)(\mathbf{x}_k - \mathbf{a}_k) + \mathcal{R}_1(\|\mathbf{x}\|^3),\end{aligned} \tag{19}$$

where the Jacobian is $\mathbf{J}_i^j(\mathbf{x}) = \nabla_i^j \mathbf{f}(\mathbf{x})$ is an order 2 tensor of dimension $n \times 6$, the Hessian tensor is an order 3 tensor $\mathbf{H}_i^{jk}(\mathbf{x}) = \nabla_i^{jk}\mathbf{f}(\mathbf{x})$ of dimension $n \times 6 \times 6$ and $\mathcal{R}_1(\|\mathbf{x}\|^3)$ is the third order remainder. The Jacobian $\mathbf{J}$ can also be approximated via a Taylor series expansion as:

$$\mathbf{J}_i^j(\mathbf{x}) = \mathbf{J}_i^j(\mathbf{a}) + \mathbf{H}_i^{jk}(\mathbf{a})(\mathbf{x}_k - \mathbf{a}_k) + \mathcal{R}_2(\|\mathbf{x}\|^2). \tag{20}$$

Substituting for $\mathbf{H}_i^{jk}(\mathbf{a})(\mathbf{x}_k - \mathbf{a}_k)$ from (20) into (19) and evaluating at $\mathbf{a} = \mathbf{0}$ gives:

$$\mathbf{f}_i(\mathbf{x}) = \mathbf{f}_i(\mathbf{0}) + \frac{1}{2}\left(\mathbf{J}_i^j(\mathbf{0}) + \mathbf{J}_i^j(\mathbf{x})\right)\mathbf{x}_j + \mathcal{R}_3(\|\mathbf{x}^3\|). \tag{21}$$

It can be seen here that if the third order terms are ignored, the second order approximation depends on both the Jacobian evaluated at the current position $\mathbf{J}(\mathbf{0})$ and the Jacobian evaluated at the solution $\mathbf{J}(\mathbf{x})$. Since $\mathbf{x}$ is the unknown solution to the system of equations, it would seem impossible to determine this term. However, it will be shown that in the present tracking case, it is possible to substitute the current image in (17) for the reference image (image at the solution) in order to obtain an equivalent term without knowing $\mathbf{x}$.

In the case of the stereo image function, the second order expansion is then given as:

$$\boldsymbol{\mathcal{I}}(\tilde{\mathbf{x}}) \approx \boldsymbol{\mathcal{I}}(\mathbf{0}) + \frac{\mathbf{J}(\mathbf{0}) + \mathbf{J}(\tilde{\mathbf{x}})}{2}\tilde{\mathbf{x}}, \tag{22}$$

where $\mathbf{J}(\mathbf{0})$ is the current image Jacobian and $\mathbf{J}(\tilde{\mathbf{x}})$ is reference image Jacobian.

The current Jacobian and the reference Jacobians can be decomposed as the product of four Jacobians. Their detailed derivation can be found in Appendix IX-A. In summary of the discussion found in the appendix, the second order approximation of equation (22) gives:

$$\mathcal{J}(\tilde{\mathbf{x}}) = \frac{(\mathbf{J}_{\boldsymbol{\mathcal{I}}} + \mathbf{J}_{\boldsymbol{\mathcal{I}}^*})}{2}\mathbf{J}_{\mathbf{w}}\mathbf{J}_{\mathbf{T}}\mathbf{J}_{\mathcal{V}}, \tag{23}$$

where only $\mathbf{J}_{\boldsymbol{\mathcal{I}}}$ varies with time and needs to be computed at each iteration.

The objective function is minimised by iteratively solving (17) by using (23) and (14) for:

$$\tilde{\mathbf{x}} = -\lambda\left(\mathbf{D}\mathcal{J}\right)^{+}\mathbf{D}(\boldsymbol{\mathcal{I}} - \boldsymbol{\mathcal{I}}^*), \tag{24}$$

where $(\mathbf{D}\mathcal{J})^+$ is the pseudo-inverse, $\mathbf{D}$ is a diagonal weighting matrix determined from a robust function (see Appendix IX-C) and $\lambda$ is the gain which ensures an exponential decrease of the error. Refer to Figure 3 for a summary of this estimation process.

## V. IMPLEMENTATION

### A. Multi-resolution tracking

In order to improve the computational efficiency of the approach and to handle large displacements a multi-resolution reference image [27], [28] was constructed and used for tracking (refer to Figure 4). As is commonly done in this type of approach, the tracking begins at the highest levels (the lowest resolution) and performs tracking at this level until convergence. The $i^{\text{th}}$ resolution image is obtained by simply warping the original images as:

$$\boldsymbol{\mathcal{I}}_i = \boldsymbol{\mathcal{I}}(w_H(\boldsymbol{\mathcal{P}}, \mathbf{H}_i)), \tag{25}$$

where $w_H$ is a stereo homographic warping function that warps the points $\boldsymbol{\mathcal{P}}$ using the $3 \times 3$ homography $\mathbf{H}_i$ as:

$$\mathbf{p}_i = \begin{bmatrix} scaleu_i & 0 & 0 \\ 0 & scalev_i & 0 \\ 0 & 0 & 1 \end{bmatrix}\mathbf{p} \tag{26}$$

Note that this is equivalent to changing the calibration parameters in the stereo warping function of (14).

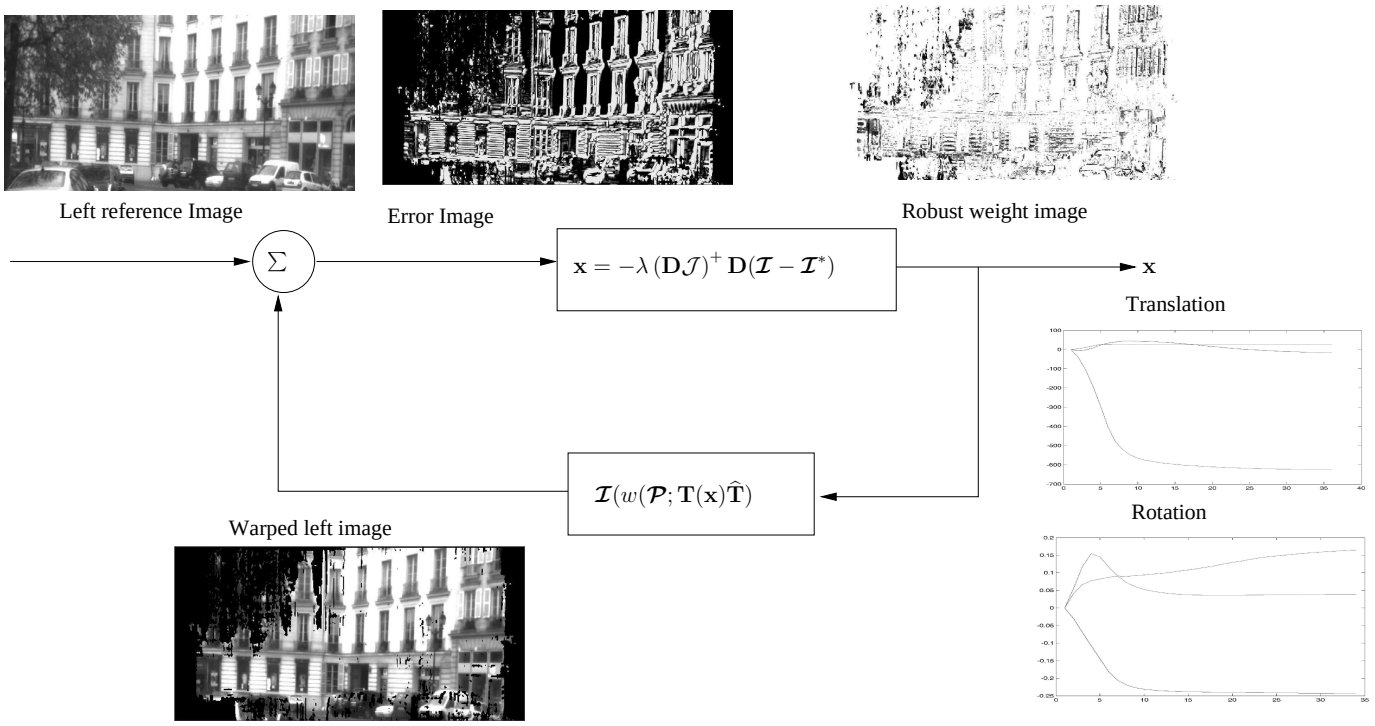



Fig. 3. The iterative estimation process as given by equation (24). The reference image is given as input. An error image is then obtained with the warped current image. The robust second order term is inversed to obtain an incremental pose (parametrised as an element of the Lie algebra here). The current image is then re-warped and the process is repeated until convergence. Only the left images and corresponding data are given here, however, the same is available for the right image.

It should be noted here that the Nyquist-Shannon image re-sampling theorem is a widely studied problem in image processing [29], [27] and computer graphics [30] and it is well known that aliasing effects can occur if this is not done correctly. In general Gaussian smoothing is performed to eliminate aliasing but here we use simple bi-linear interpolation which acts as a local box filter that provides a reasonable trade-off between computationally efficient filtering and aliasing.

Once convergence is obtained the current pose is used to initialise the next level in the pyramid and tracking is once again performed until convergence. This is repeated until the highest resolution of the pyramid is reached. In this way the larger displacements are minimised at lower cost on smaller images. Furthermore, the smaller images smooth out much of the detail required for fine adjustment and provide more global information that helps target the larger movements initially. This also has the effect of avoiding certain local minima.

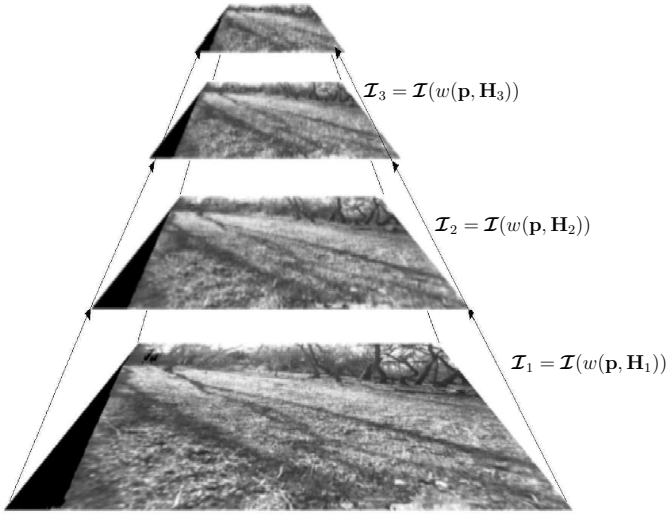



Fig. 4. A multi-resolution pyramid used to improve the computational performance of the tracking and avoid local minima. The homographic transform between scales is given in Figure 25

### *B. Reference Image-pairs*

Depending on the application, the stereo 3D visual odometry approach can acquire an image-based model

online via correspondences (see Section V-C) and the odometer does not require initialisation with respect to the world coordinate system. On the other hand, if it is assumed that a reference sequence is available together with pre-matched disparities, then initialisation must be performed by finding the closest images and then performing tracking to initialise the pose. This problem has not treated here, however, we did perform tracking at the loop closure showing that the convergence domain is quite large (see Section VI).

As the camera pair moves through the scene the reference image may be no longer visible or the warped resolution becomes so poor that it is necessary to interpolate many pixels. In both cases this leads to miss-tracking. Therefore, in order to perform large scale tracking it is necessary to continually update the reference image pair $\boldsymbol{\mathcal{I}}^*$. An update is detected by monitoring the error norm along with a robust estimate of the scale of the error distribution (i.e. the Median Absolute Deviation). As soon as they become too large another set of dense correspondences between the stereo pair is made so as to reinitialise the tracking. As long as the same reference image is used then the minimisation cut-off thresholds can be tuned for speed since the next estimation will recover any remaining error, however, if the reference image is changed the previous estimate is minimised with smaller cut-off thresholds so as to minimise any drift that may be left over.

### C. Dense Correspondences

As mentioned, the reference image pair(s) need to be initialized with dense correspondences. The correspondence problem has been heavily studied in the computer vision literature and many different approaches are possible [31]. When the cameras are calibrated the correspondence problem reduces to a 1D search along epipolar lines. This can either be performed off-line in a learning phase or on-line at each new acquisition depending on computational requirements (real-time approaches are feasible [32]). In this paper the approaches given in [31], [33] along with a custom real-time GPU implementation for basic SSD correlation were used and tested. Nevertheless, any other type of dense correspondence algorithm could be used. The method given in [33] was initially used since it is particularly suited to urban canyon environments since the notions of horizontal and vertical slant are used to approximate first-order piecewise continuity. In this way the geometric projection of slanted surfaces from $N$ pixels on one epipolar line to $M$ pixels on another is not necessarily one-to-one but can be many-to-one or one-to-many. See Figure 5(a) for correspondence results of a typical image pair. In this case the disparity search region was fixed in a range of $-20$ to $-180$ pixels along the epipolar lines. In figure 5(b) the visual quality of the results can be inspected. In this case the points in the right reference image are warped to the left by interpolation and it can be seen that the results is quite similar to the original image in 5(c) apart from the fact that there are occluded regions in black where the dense matching algorithm did not succeed.

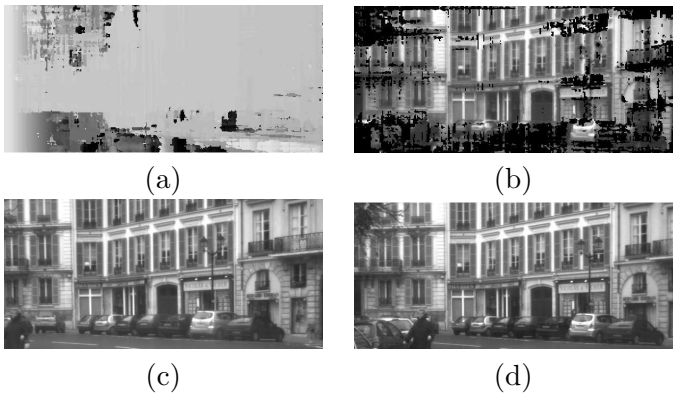
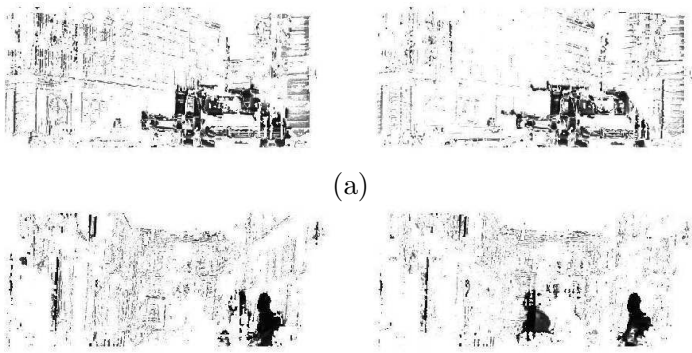



Fig. 5. Dense correspondence of an urban canyon with correspondence occlusion in black: (a) the disparity image from left to right and (b) right image warped to the left image using the disparity values with occluded disparities in black, (c,d) right and left original images respectively.

### D. Robust Estimation

A robust M-estimation technique (as detailed in Appendix IX-C and [26]) was used to reject outliers not corresponding to the definition of the objective function. The use of robust techniques is very interesting in the case of a highly redundant set of measurement as is the case of a set of dense correspondences. The outliers generally correspond to occlusions, illumination changes, matching error, noise in the image or the self occlusion of the corners of the buildings.

In figure 6 (a) a moving truck has been rejected as an outlier whilst a stationary truck in the background was used to estimate the pose. In this way it can be seen that the proposed algorithm has exploited all the useful information in the image so as to estimate the pose. In figure 6 (b) a moving pedestrian has been rejected and it can be seen that both the pedestrian projected from the reference image as well as the current position of the pedestrian have been rejected. This type of information could be useful in an application for determining the trajectory of moving



(a)

(b)

Fig. 6. Robust outlier rejection: Two images showing the outlier rejection weights. The darker the points, the less influence they have on the estimation process. In (a) it can be seen that a moving truck has been rejected. In (b) a moving pedestrian has been rejected. It can also be noted that other outliers are detected in the image. These points generally correspond to matching error, noise in the image or the self occlusion of the corners of the buildings.

obstacles.

### *E. Extended Kalman Filter*

In the context of tracking the trajectory of a car, very large inter-frame movements are observed. In the sequences considered in the following results, typical inter-frame movement was 1-2 meters per image with a car travelling between 50 and 70 km/hr. Even though tracking succeeds without predictive filtering, in order to improve computational efficiency (significantly less iterations in the minimisation) it was necessary to implement a predictive filter. In this paper the well known Extended Kalman filter described in [34] was used for filtering the pose (i.e. the pose estimate was considered to be the measurement input to the filter).

## VI. Results

### *A. Simulation results*

In order to test the algorithm with a ground truth a synthetic video sequence was created by warping real images onto various 3D surfaces. In this way realistic images were created with a known ground truth about the trajectory of the camera. In order to only test the capability of the proposed tracker the true image correspondences were provided.

In Figure 7, a 3D sphere is considered. A patch is selected on the sphere and it can be seen that the contour of the patch warps correctly with the contour of the sphere throughout the tracking process. In the final images of the sequence (c) and (d), the pixels which are on the edge of the sphere begin to become occluded. In this case the rigid geometric structure defined within the quadrifocal estimation naturally wraps the edge of the sphere back onto itself. i.e. this is only feasible geometric solution to the estimation problem. Of course, if a robust estimator is not used then tracking fails but with the M-estimator these pixels are no longer used for estimation and are rejected. A test was, however, performed when there was no occlusion which showed that the robust estimator required 10% more iterations to converge than without (or was not as precise with the same number of iterations). There is therefore a compromise to be made between efficiency, precision and robustness. In (e) one can see the estimated trajectory of the camera pair. In (f) it can be seen that even if the RMS error is very small, there is an increase in interpolation error as the camera gets further away from the reference image.

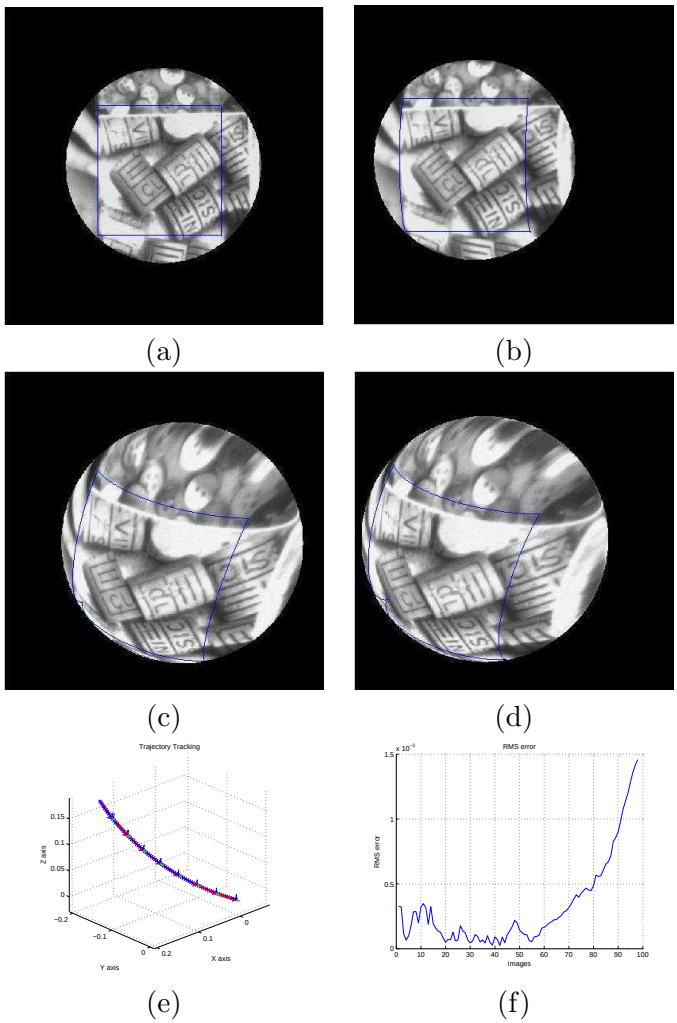



Fig. 7. Simulation of tracking a patch (outlined by the contour) on a 3Dsphere: (a,b) Initial left and right images, (c,d) final left and right images, (e) the estimated trajectory of the sphere, (f) the RMS error between the estimated trajectory and the true one showing drift due to distance from the reference image.

### B. Visual Odometry on Real Sequences

The algorithm was first tested on several real full-scale sequences of urbain scenes from different streets in Versailles, France, as can be seen in Figure 8 and 9. Radial distortion has been removed from the images before processing. In these sequences the stereo images are of size $760 \times 578$, the cameras are not rectified and have a baseline of approximately $1.01m$. In the experiments the bottom half of the image was remove since it contains only the road. These and the following video demonstrations plus more are available online at the authors websites.

The sequence shown in Figure 8 is that of a relatively straight road. The distance travelled by the car has been measured using road markings in the images and satellite views with a precision of $2.9cm/pixel$ for


Fig. 8. Trajectory tracking along a road in Versailles : (a) The trajectory shown in white has been superimposed on a satellite image. An typical stereo image is shown at the top.

the Versailles region. The path length measured by both Google earth and the tracker was about $440m$. It is difficult to register the satellite image with the projection of the trajectory since no three non-collinear points were available and the best that can be said is that they are approximately the same absolute length (ignoring tilt of the cameras and the incline of the road). Throughout the sequence several moving vehicles pass in front of the cameras and at one stage a car is overtaken.

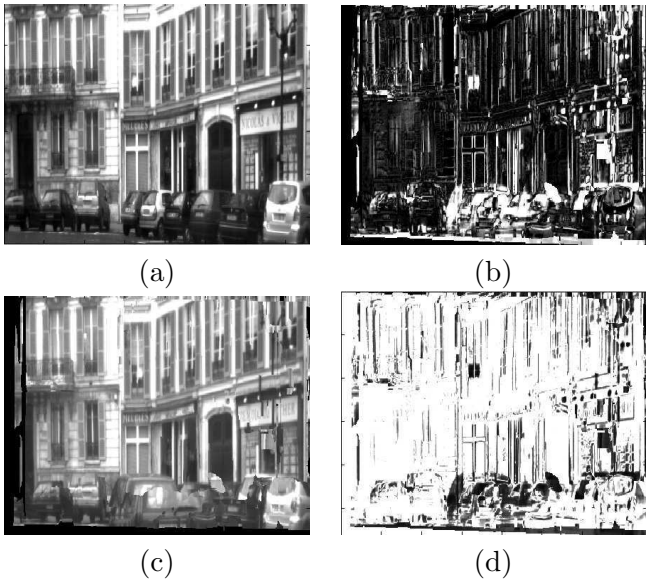
The sequence shown in Figure 9, is particularly illustrative since a full loop of the round-about was performed. In particular this enables the drift to be measured at the crossing point in the trajectory. The drift was measured by comparing the integrated odometry from the trajectory going around the round about with the pose measured between images that are close at the closure of the loop (see Figure 10). Although there is no real loop intersection, it is clear that a single pose estimate is more accurate that integrating a long trajectory. Nevertheless, there is a non-negligible distance between the images which means that a sufficient overlap is required between the views. By performing this experiment it is also possible to give an idea of the convergence domain for the proposed technique (even if this depends highly on the scene being viewed). For the loop closing from the $22^{\text{nd}}$ to the

$471^{\text{st}}$ images the drift estimate was

$$T_{drift} = \text{inv}(\text{T}_{22} * \text{T}_{\text{loopclose}}) * \text{T}_{471} \qquad (27)$$

which corresponds to the pose:
$(-81.0cm, -60.8cm, 92.9cm, 0.1^o, 0.2^o, 0.2^o)$.
This results gives an RMS error at the crossing point of 137.5cm leading to a drift of approximately 8.6% drift.

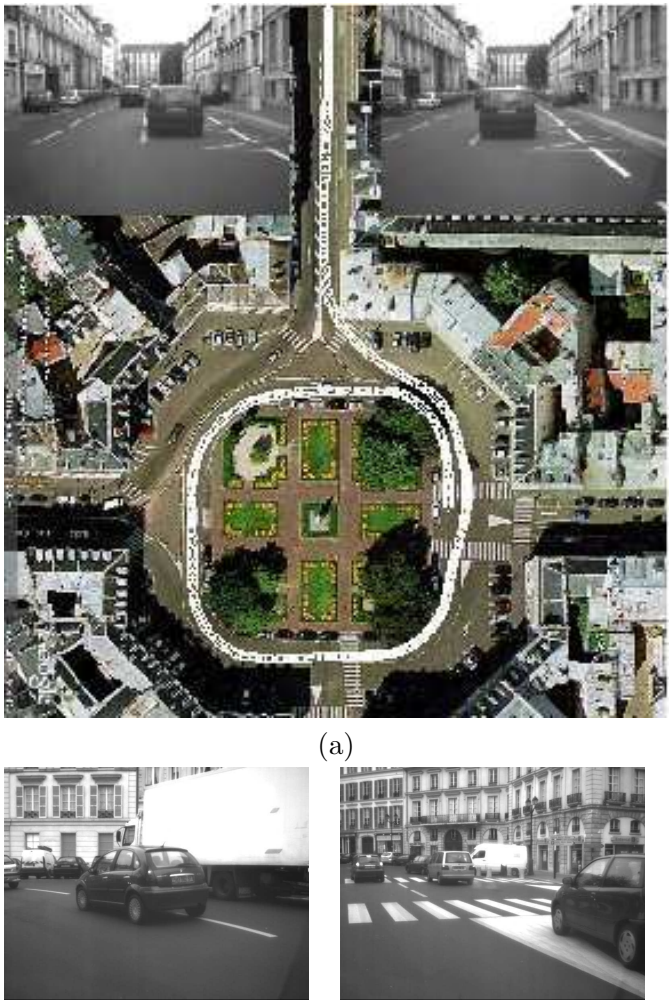
In the case of large scale scenes such as this one it was necessary to detect and update the reference image periodically when it was no longer visible or too approximate. This update was detected by putting a maximum threshold on the error norm along with a threshold on the Median Absolute Deviation (a robust measure of the standard deviation) and re-initializing the dense correspondences. Due to the highly redundant amount of data, the robust estimator was able to successfully reject pedestrians and moving cars from the estimation process. It can be noted, however, that all static information available was used to estimate the pose (including the parked cars) therefore leading to a very precise result with minimal drift over large displacements.



(a) (b) (c) (d)

Fig. 10. Loop closing between the $22^{\text{nd}}$ and the $471^{\text{st}}$ images of the Versailles round about sequence. (a) The left reference image (the $471^{\text{st}}$ image), (b) The final image error for the left camera, (c) The $22^{\text{nd}}$ image warped to the position of the reference image, (d) The estimated rejection weights. It can be seen that there is a global change in illumination between the images leading to a higher final error and more outliers than in the incremental tracking case.

The experiment shown in Figure 11 is that of a robotic airship with a wide baseline stereo pair looking down towards the ground. A full loop is performed around a parking lot. This application demonstrates the full potential of the approach since the full 6dof visual odometry is necessary in the aerial domain so as to provide the position of the airship in 3D space. It is important to note that due to the fact that the cameras are facing downwards, the movement observed within the images is very rapid and rough. Furthermore, the balloon is also quite unstable and subject to wind variation. Consequently, the use of predictive



(a) (b) (c)

Fig. 9. Trajectory tracking around a round-about in Versailles : (a) The trajectory shown in white has been superimposed on a satellite image and it can be seen visually that the trajectory aligns with the four corners of the round-about (4 points are required to estimate the pose). The length of the path is approximately 392m taken in 698 images. The maximum inter-frame displacement was $1.78m$ and the maximum inter-frame rotation was $2.23^o$ (b and c) several occlusions which occurred during the sequence and image 300 and 366 respectively (on the right side of the round-about).

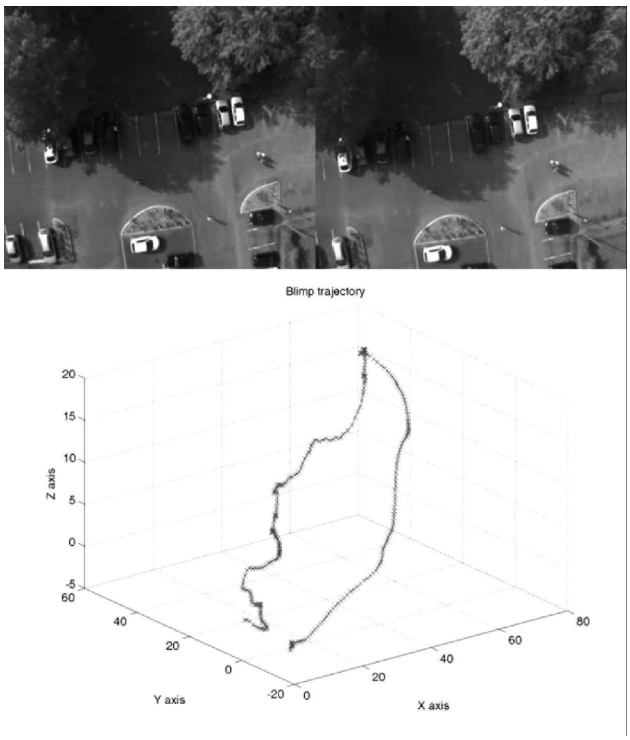



Fig. 11. Trajectory tracking from the LAAS blimp sequence: (a) The trajectory is shown underneath and a pair of typical stereo images is given above.

filtering in such a context is most often worse than not having a filter. The results in Figure 11 have therefore been obtained without any filter. Unfortunately, this means that there are large inter-frame movements and the visual odometry requires some time to converge (many more iterations). Nevertheless, with the combined use of multi-resolution tracking and the selection of the strongest image gradients the visual odometry computation is still able to be maintained in real-time. The results given here have been obtained with an image resolution divided by 6 and only the 50,000 strongest gradients have been used. Finally, it can also be noted that some local minima were observed at full resolution when the car parking spaces were displaced by the movement of the camera such that the current image of the cars where shifted into an adjacent car park with respect to the reference image. This is due to the fact that the symmetry of the visual information corresponds to the movement of the camera. This problem, however, was also avoided by considering a different resolution in the pyramid.

In Figure 12 another experiment is given for a large outdoor sequence with rough natural terrain. Here the images are rectified of size $512 \times 384$ and captured at about 10Hz. In this case RTK GPS ground truth data was available that is accurate to within several cm. It can be seen here that there is a systematic drift in the direction normal to the earth surface along with drift that depends on the direction of rotation. This could be due to various factors including camera calibration errors (extrinsic or intrinsic) or errors in the dense matching between the left and the right images. Towards the end of the trajectory a large deviation can be seen due to a very large rotation on the last corner leading to a tracking failure. This could be dealt with by integrating inertial sensors into the framework as has been done in [13].

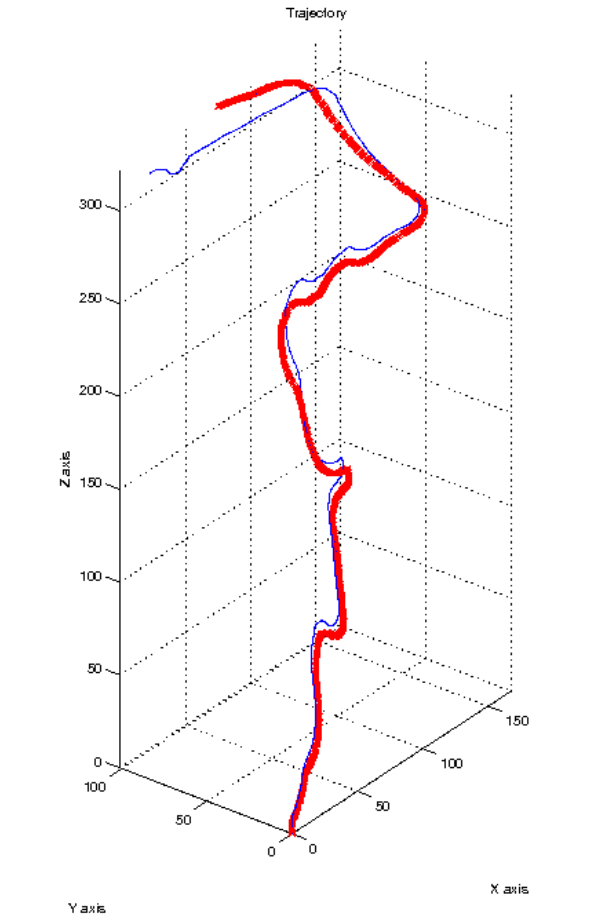
### C. Computational requirements

An optimised version has been implemented which is capable of running in real-time (30Hz) on a laptop Dual Core system for images of size $100 \times 100$. Of course this computation time varies depending on the size of the image used, the precision required and the magnitude of the displacements considered. Furthermore, a dense correspondence algorithm is also used that runs in real time on the GPU. With dual core parallelization of the left and right images along with a sub pixel dense correspondence algorithm running in parallel on the GPU, a full image of size $759 \times 560$ requires only $500ms$ to perform visual odometry. Although the robustness is reduced by using less information, it can be noted that there is only a small difference in precision between the full and reduced images. With the numbers given here there was only about 0.004% drift in translation and $0.03deg/deg$ drift in rotation when measured from a 360m long sequence. The real-time implementation of this approach has required much optimisation that will not be detailed here, however, the various approaches include in terms of dense correspondence:

*1.* Use an off-line training sequence to obtain a set of dense corresponding reference image-pairs for use online.
*2.* Perform dense correspondences online (i.e. GPU).

In terms of the visual odometry:

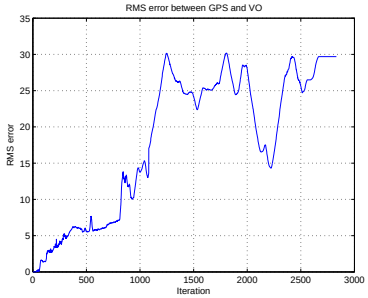
*1.* Contract the quadrifocal transfer function into constant and non-constant components.
*2.* Decompose the Jacobian as outlined in Annex IX-A.
*3.* Parallelize the computation for SIMD architectures such as multi-core and GPU processors.
*4.* Improve the algorithms by multi-resolution and feature selection (i.e. the strongest gradients).




(a)

(b)

(c)




(d)

Fig. 12. The outdoor Dunes sequence taken from a ground vehicle with (a) GPS ground truth data in blue and the visual odometry estimation in red. (b) and (c) show two images from this difficult sequence. (d) The RMS error in meters between the GPS and visual sensors, it can be seen that there is significantly more drift when there are large rotations at iterations anti-clockwise from 1081 to 1244 and clockwise from 1996 to 2121.

## VII. Conclusions and Future Works

The quadrifocal tracking methodology described in this paper has shown to be very efficient, accurate (very small drift) and robust over a wide range of scenarios. The approach is very interesting because trajectory estimation is integrated into a single global sensor-based process that does not depend of intermediate level feature extraction and matching. Tracking is initialised automatically at the origin within the visual odometry approach. Furthermore, a compact image-based stereo model of the environment may be obtained using standard dense stereo correspondence algorithms and instead of explicit estimation of an a-priori 3D model. The robust efficient second order minimisation technique also allows minimisation of a highly redundant non-linear function in a precise manner. Indeed the algorithm rejects outliers such as pedestrians, traffic, building occlusions and matching error.

Further work will be devoted to estimating optimal stereo image-based models of the environment by updating the dense correspondences in a Simultaneous Localisation and Correspondence style approach. It would be interesting to test loop closing procedures and devise strategies to recognise previously seen places within this framework.

## VIII. Acknowledgements

This study was part of the French national MOBIVIP PREDIT project aimed at autonomous vehicle navigation in urban environments.

## IX. Appendices

### A. Appendix I: Jacobian computation

#### A.1 Current Jacobian

The current Jacobian $\mathbf{J}(\mathbf{0})$ can be obtained by taking the derivative of (5) and evaluating at $\mathbf{x} = \mathbf{0}$ as:

$$\mathbf{J}(\mathbf{0}) = \left[\nabla_{\mathbf{x}} \boldsymbol{\mathcal{I}} \left( w \left( \boldsymbol{\mathcal{P}}^{*}, \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}) \right) \right) \right]_{\mathbf{x}=\mathbf{0}} \tag{28}$$

Taking into account property (16) gives:

$$\mathbf{J}(\mathbf{0}) = \left[\nabla_{\mathbf{x}} \boldsymbol{\mathcal{I}} \left( w \left( w \left( \boldsymbol{\mathcal{P}}^{*}, \mathbf{T}(\mathbf{x}) \right) \right), \widehat{\mathbf{T}} \right) \right]_{\mathbf{x}=\mathbf{0}}, \tag{29}$$

which can be written as a product of four Jacobians:

$$\mathbf{J}(\mathbf{0}) = \mathbf{J}_{\boldsymbol{\mathcal{I}}} \mathbf{J}_{w} \mathbf{J}_{\mathbf{T}} \mathbf{J}_{\mathcal{V}} \tag{30}$$

**1.** $\mathbf{J}_{\boldsymbol{\mathcal{I}}}$ is of dimension $1 \times 2 * 3 = 6$ and corresponds to the spatial derivative of the pixel intensities for

each of the current images warped by the projective transformation $w(\mathbf{z}, \widehat{\mathbf{T}})$:

$$\mathbf{J}_{\boldsymbol{\mathcal{I}}} = \left[\nabla_{\mathbf{z}} \boldsymbol{\mathcal{I}} \left( w(\mathbf{z}, \widehat{\mathbf{T}}) \right)\right]_{\mathbf{z}=\boldsymbol{\mathcal{P}}^*} . \tag{31}$$

**2**. $\mathbf{J}_w$ is of dimension $6 \times 2 * 16 = 32$ and corresponds to the derivative of the pixel location with respect to the elements of two homogeneous pose matrices embedded within each of the two trifocal tensor of the warping function:

$$\mathbf{J}_w = \left[\nabla_{\mathcal{Z}} w\left(\boldsymbol{\mathcal{P}}^*, \mathcal{Z}\right)\right]_{\mathcal{Z}=\mathbf{T}(\mathbf{0})=\mathbf{I}} \tag{32}$$

**3**. $\mathbf{J}_{\mathbf{T}}$ is of dimension $32 \times 6 * 2 = 12$ and can be written as:

$$\mathbf{J}_{\mathbf{T}} = \nabla_{\mathbf{x_L}, \mathbf{x_R}} \left[ \begin{array}{c} \mathbf{T_L}(\mathbf{x_L}) \\ \mathbf{T_R}(\mathbf{x_R}) \end{array} \right]_{\mathbf{x_L}=\mathbf{x_R}=0} . \tag{33}$$

where both left and right camera matrices are not expressed in the same reference frame but with respect to their canonical trifocal tensor bases. This Jacobian can be written as:

$$\mathbf{J}_{\mathbf{T}} = [\mathbf{a_L}_1, \ldots, \mathbf{a_L}_6, \mathbf{a_R}_1, \ldots, \mathbf{a_R}_6] \tag{34}$$

where the vectors $\mathbf{a_i}$ are obtained by reshaping the generator matrices $A_i$ (columns then rows). The generators are the basis of the Lie algebra $\mathfrak{se}$ and can be determined from equation (4) by selecting a basis $x$ from the twist $\mathbf{x}_i = (\boldsymbol{\upsilon}_i, \boldsymbol{\omega}_i)$ (i.e. $\mathbf{v_1} = (1, 0, 0, 0, 0, 0)$) to obtain:

$$\mathbf{A}_i = \left[ \begin{array}{cc} [\boldsymbol{\omega}_i]_\times & \boldsymbol{\upsilon}_i \\ \mathbf{0} & \mathbf{0} \end{array} \right] \tag{35}$$

**4**. $\mathbf{J}_{\mathcal{V}}$ of dimension $12 \times 6$ is a pair of adjoint maps that transform the twists from the precedent Jacobian into the same reference frame according to (12). It is chosen to center the two components of $\mathbf{J}_{\mathbf{T}}$, corresponding to the left and right canonical coordinate systems, so that they represent the same minimal set of unknown parameters. This corresponds to the application of the bilinear constraint to the pair of trifocal constraints so as to form a quadrilinear one giving:

$$\mathbf{J}_{\mathcal{V}} = \left[ \begin{array}{c} \frac{\partial \mathbf{x_L}}{\partial \mathbf{x}} \\ \frac{\partial \mathbf{x_R}}{\partial \mathbf{x}} \end{array} \right] = \left[ \begin{array}{c} \mathbf{V_L} \\ \mathbf{V_R} \end{array} \right] \tag{36}$$

where the adjoint map is given as:

$$\mathcal{V} = \left[ \begin{array}{cc} \mathbf{R^c} & \mathbf{t^c}_\times \mathbf{R^c} \\ \mathbf{0}_3 & \mathbf{R^c} \end{array} \right], \tag{37}$$

and where $\mathbf{T^c} = (\mathbf{R^c}, \mathbf{t^c})$ is the centering pose given in (11), which maps the current left camera matrix to the stereo center according to (11). Similarly, an adjoint map can be obtained to transform the twist of the right current camera with respect to the right reference camera using $\mathbf{T^{c\prime}}$.

Three of the Jacobians, $\mathbf{J}_w$, $\mathbf{J}_{\mathbf{T}}$ and $\mathbf{J}_{\mathcal{V}}$ are constant and need only be calculated once for the reference image. The Jacobian $\mathbf{J}_{\boldsymbol{\mathcal{I}}}$ must be calculated at each iteration.

### A.2 Reference Jacobian

The reference Jacobian $\mathbf{J}(\tilde{\mathbf{x}})$ can be obtained by taking the derivative of (5) as:

$$\mathbf{J}(\mathbf{x}) = \left[\nabla_{\mathbf{x}} \boldsymbol{\mathcal{I}} \left( w(\boldsymbol{\mathcal{P}}^*, \widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})) \right)\right]_{\mathbf{x}=\tilde{\mathbf{x}}} \tag{38}$$

The group property (16) is reused again here to replace the current image by the reference image. This is achieved by introducing a *true* transformation $\overline{\mathbf{T}}$ corresponding to the solution that is sought after along with its inverse. In this case (38) can be rewritten as:

$$\begin{array}{rl} \mathbf{J}(\mathbf{x}) = & \left[\nabla_{\mathbf{x}} \boldsymbol{\mathcal{I}} \left( w(w(\boldsymbol{\mathcal{P}}^*, \overline{\mathbf{T}}^{-1}\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})) \right), \overline{\mathbf{T}}) \right]_{\mathbf{x}=\tilde{\mathbf{x}}} \\ & \left[\nabla_{\mathbf{x}} \boldsymbol{\mathcal{I}}^* \left( w(\boldsymbol{\mathcal{P}}^*, \overline{\mathbf{T}}^{-1}\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})) \right)\right]_{\mathbf{x}=\tilde{\mathbf{x}}} \end{array} \tag{39}$$

The reference Jacobian can now be written as a product of four Jacobians:

$$\mathbf{J}(\tilde{\mathbf{x}}) = \mathbf{J}_{\boldsymbol{\mathcal{I}}^*} \mathbf{J}_{w^*} \mathbf{J}_{\mathbf{T}^*} \mathbf{J}_{\mathbf{V}^*} \tag{40}$$

**1**. $\mathbf{J}_{\boldsymbol{\mathcal{I}}^*}$ is of dimension $1 \times 2 * 3 = 6$ and corresponds to the spatial derivative of the pixel intensities for each of the reference images warped by the projective transformation $w(\mathbf{z}, \overline{\mathbf{T}}^{-1}\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}))$:

$$\mathbf{J}_{\boldsymbol{\mathcal{I}}^*} = \left[\nabla_{\mathbf{z}} \boldsymbol{\mathcal{I}}^* \left(w(\mathbf{z}, \mathbf{I})\right)\right]_{\mathbf{z}=\boldsymbol{\mathcal{P}}^*} . \tag{41}$$

where $\overline{\mathbf{T}}^{-1}\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}) = \mathbf{I}$ at $\mathbf{x} = \tilde{\mathbf{x}}$.

**2**. $\mathbf{J}_{w^*}$ is of dimension $6 \times 2 * 16 = 32$ and corresponds to the derivative of the pixel location with respect to the elements of two homogeneous pose matrices embedded within each of the two trifocal tensor of the warping function:

$$\mathbf{J}_{w^*} = \left[\nabla_{\mathcal{Z}} w\left(\boldsymbol{\mathcal{P}}^*, \mathcal{Z}\right)\right]_{\mathcal{Z}=\mathbf{I}} = \mathbf{J}_w \tag{42}$$

**3**. $\mathbf{J}_{\mathbf{T}^*}$ depends on the unknown twist $\tilde{\mathbf{x}}$ which is the solution to the estimation problem. However, by using the group properties it can be shown that the following proposition is true (see Appendix II):

$$\mathbf{J}_{\mathbf{T}^*} \mathbf{J}_{\mathcal{V}^*} \mathbf{x} = \mathbf{J}_{\mathbf{T}} \mathbf{J}_{\mathcal{V}} \mathbf{x}. \tag{43}$$

**4**. $\mathbf{J}_{\mathbf{V}^*} = \mathbf{J}_{\mathcal{V}}$ since the same rigid transformation is made.

### B. Appendix II

Here a small proof is given for the proposition in (43). The proposition is first rewritten as:

$$\left[\frac{d(\overline{\mathbf{T}}^{-1}\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}))}{d\mathbf{x}}\right]_{\mathbf{x}=\tilde{\mathbf{x}}}\tilde{\mathbf{x}} = \left[\frac{d\mathbf{T}(\mathbf{x})}{d\mathbf{x}}\right]_{\mathbf{x}=\mathbf{0}}\tilde{\mathbf{x}} \qquad (44)$$

Considering the left hand side, it is possible to make a substitution of variables for $\mathbf{x} = \tilde{\mathbf{x}} + \mathbf{y}$ to give:

$$\mathbf{J}_{\mathbf{T}^*} = \left[\frac{d(\overline{\mathbf{T}}^{-1}\widehat{\mathbf{T}}\mathbf{T}(\tilde{\mathbf{x}}+\mathbf{y}))}{d\mathbf{y}}\right]_{\mathbf{x}=\mathbf{0}}\frac{d\mathbf{y}}{d\mathbf{x}}\tilde{\mathbf{x}} \qquad (45)$$

where using the group properties $\mathbf{T}(\tilde{\mathbf{x}})\mathbf{T}(\mathbf{y}) = \mathbf{T}(\tilde{\mathbf{x}}+\mathbf{y})$ and assuming that the 'true' pose $\overline{\mathbf{T}} \approx \widehat{\mathbf{T}}\mathbf{T}(\tilde{\mathbf{x}})$ gives:

$$\mathbf{J}_{\mathbf{T}^*} = \left[\frac{d\mathbf{T}(\mathbf{y})}{d\mathbf{y}}\right]_{\mathbf{y}=\mathbf{0}}\tilde{\mathbf{x}} \qquad (46)$$

which gives the same $\tilde{\mathbf{x}}$ as in the right hand side of (44) for all $\mathbf{y} = \mathbf{x} - \tilde{\mathbf{x}}$.

### C. Appendix III

This section gives a brief overview for the calculation of weights for each image feature. The weights $w_i$, which represent the different elements of the $\mathbf{D}$ matrix and reflect the confidence of each feature, are usually given by [25]:

$$w_i = \frac{\psi(\delta_i/\sigma)}{\delta_i/\sigma}, \qquad (47)$$

where $\psi\Big(\delta_i/\sigma\Big) = \frac{\partial\rho\left(\delta_i/\sigma\right)}{\partial\mathbf{r}}$ ($\psi$ is the influence function) and $\delta_i$ is the normalized residual given by $\delta_i = \Delta_i - Med(\Delta)$ (where $Med(\Delta)$ is the median operator).

Of the various loss and corresponding influence functions that exist in the literature Tukey's hard re-descending function is considered. Tukey's function completely rejects outliers and gives them a zero weight. This is of interest in tracking applications so that a detected outlier has no effect on the virtual camera motion and does not cost computational effort uselessly. This influence function is given by:

$$\psi(u) = \begin{cases} u(C^2-u^2)^2 & , \text{ if } |u| \leq C \\ 0 & , \text{else,} \end{cases} \qquad (48)$$

where the proportionality factor for Tukey's function is $C = 4.6851$ and represents 95% efficiency in the case of Gaussian Noise.

In order to obtain a robust objective function, a value describing the certainty of the measures is required. The scale $\sigma$ or the estimated standard deviation of the inlier data and is a critical value that can impact heavily on the efficiency of the method. This factor varies significantly during convergence, so it was estimated iteratively using the Median Absolute Deviation (MAD):

$$\widehat{\sigma} = \frac{1}{\Phi^{-1}(0.75)}Med_i(|\delta_i - Med_j(\delta_j)|). \qquad (49)$$

where $\Phi()$ is the cumulative normal distribution function and $\frac{1}{\Phi^{-1}(0.75)} = 1.48$ represents one standard deviation of the normal distribution.

The introduction of the weighting matrix $\mathbf{D}$ into the minimization scheme in Section IV is achieved via and iteratively re-weighted least squares implementation. Robust weights were calculated together for each component's feature set due to incompatibilities when calculating weights directly from all the object features.

## References


[1] N. Simond and P. Rives, "What can be done with an embedded stereo-rig in urban environments?" *Robotics and Autonomous Systems*, vol. 56, pp. 777–789, 2008.

[2] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd, "Real-time localization and 3D reconstruction," in *IEEE Conference of Vision and Pattern Recognition*, New-York, USA, June 2006.

[3] A. Comport, E. Marchand, M. Pressigout, and F. Chaumette, "Real-time markerless tracking for augmented reality: the virtual visual servoing framework," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 4, pp. 615–628, July 2006.

[4] A. J. Davison and D. W. Murray, "Simultaneous localisation and map-building using active vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 2002.

[5] A. Chiuso, P. Favaro, H. Jin, and S. Soatto, "Structure from motion causally integrated over time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 523–535, 2002.

[6] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, Washington, DC, USA, July 2004, pp. 652–659.

[7] R. Hartley and A. Zisserman, *Multiple View Geometry in computer vision.* Cambridge University Press, 2001, book.

[8] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry for ground vehicle applications," *Journal of Field Robotics*, vol. 23, p. 2006, 2006.

[9] A. Howard, "Real-time stereo visual odometry for autonomous ground vehicles," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nice, France, 22-26 September 2008.

[10] S. Benhimane and E. Malis, "Real-time image-based tracking of planes using efficient second-order minimization," in

*IEEE International Conference on Intelligent Robots Systems*, Sendai, Japan, 28 September - 2 October 2004.

[11] G. Silveira, E. Malis, and P. Rives, "An efficient direct approach to visual slam," *IEEE Transactions on Robotics.*, vol. 20, no. 5, pp. 969–979, October 2008.

[12] S. Avidan and A. Shashua, "Threading fundamental matrices," *Pattern Analysis and Machine Intelligence*, vol. 23, no. 1, pp. 73–7, Janvier 2001.

[13] K. Konolige and M. Agrawal, "Frameslam: from bundle adjustment to realtime visual mapping," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1066–1077, October 2008.

[14] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence*, vol. 2, Vancouver, Canada, August 1981, pp. 674–679.

[15] M. Irani and P. Anandan, "About direct methods," in *In IEEE International Conference on Computer Vision: Proceedings of the International Workshop on Vision Algorithms.* London, UK: Springer-Verlag, 2000, pp. 267–277.

[16] G. Hager and P. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025–1039, October 1998.

[17] S. Baker and I. Matthews, "Equivalence and efficiency of image alignment algorithms," in *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition*, December 2001.

[18] E. Malis, "Improving vision-based control using efficient second-order minimization techniques," in *IEEE International Conference on Robotics and Automation*, vol. 2, New Orleans, April 26-May 1 2004, pp. 1843–1848.

[19] A. Comport, E. Malis, and P. Rives, "Accurate quadri-focal tracking for robust 3D visual odometry," in *IEEE International Conference on Robotics and Automation*, Rome, Italy, April 2007.

[20] B. Triggs, "The geometry of projective reconstruction. I: Matching constraints and the joint image," in *IEEE International Conference on Computer Vision*, Cambridge, MA, June 20-23 1995, pp. 338–343.

[21] O. Faugeras and B. Mourrain, "On the geometry and algebra of the point and line correspondences between n images." in *IEEE International Conference on Computer Vision*, Cambridge, MA, June 20-23 1995, pp. 951–956.

[22] R. Hartley, "Multilinear relationships between coordinates of corresponding image points and lines." in *Proceedings of the Sophus Lie Symposium*, Nordjordeid, Norway, August 1995.

[23] A. Heyden and K. Astrom, "Algebraic properties of multilinear constraints," *Mathematical Methods in the Applied Sciences*, vol. 20, no. 13, pp. 1135–1162, 1997.

[24] A. Shashua and L. Wolf, "On the structure and properties of the quadrifocal tensor," in *European Conference on Computer Vision*, 2000, pp. 710–724.

[25] P.-J. Huber, *Robust Statistics*. Wiler, New York, 1981.

[26] A. Comport, E. Marchand, and F. Chaumette, "Statistically robust 2D visual servoing," *IEEE Transactions on Robotics*, vol. 22, no. 2, pp. 415–421, April 2006.

[27] P. Burt, "The pyramid as structure for efficient computation," *Multiresolution Image Processing and Analysis, SpringerVerlag*, pp. 6–35, 1984.

[28] J.-M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *Journal of Visual Communication and Image Representation*, vol. 6, no. 4, pp. 348–365, December 1995.

[29] A. Howard, "Scale-space filtering," in *8th International Joint Conference on Artificiel Intelligence*, Karlaruhe, Germany, 1983, pp. 1019 – 1022.

[30] P. Heckbert, "Fundamentals of texture mapping and image warping," Master's thesis, CS Division, U.C. Berkeley, June 1989.

[31] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision, Kauai, HI.*, December 2001.

[32] W. van der Mark and D. Gavrila, "Real-time dense stereo for intelligent vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 38–50, March 2006.

[33] A. Ogale and Y. Aloimonos, "Shape and the stereo correspondence problem," *International Journal of Computer Vision*, vol. 65, no. 1, October 2005.

[34] Z. Zhang and O. Faugeras, "Three dimensional motion computation and object segmentation in a long sequence of stereo frames," *International Journal of Computer Vision*, vol. 7, no. 3, pp. 211–241, 1992.